김소연, 김다윤, 강우원

NLP 감정분석 기반
마케팅 시장 분석

온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템

빅데이터미네이터

“ ☹😐☺

주제 :
온라인 커뮤니티 특화 감성 사전 구축을 위한 새로운 용어 극성값 분석 시스템 ”

<지난주 주요 피드백>

▪ 특허 아이디어 다시 구체화하기

→ 추출한 새로운 용어를 클러스터링 한 후, 그룹별로 극성값 부여하는 방법 구체화

구현 진행 사항

- “Kiwipiepy” 라이브러리로 사전에 없는 새로운 용어 추출

  → 전처리 후 사전 미등록 명사 추출해주는 라이브러리

  → 새로운 용어를 명사 단위로 추출하여 사용자 사전에 추가함

  ⇒ 새로운 용어 추출 결과가 부정확하고, 사전에 추가할 단어가 너무 많음

  ⇒ 사용자 사전을 학습시키면 용어 추출의 거의 안 됨

```
In [11]: inputs = list(open('natepann.txt', encoding='utf-8'))
         kiwi.extract_words(inputs, min_cnt=1, max_word_len=10, min_score=0.15, pos_score=-3.0, lm_filter=True)

Out[11]: [('서이것저것', 0.4674535095691681, 6, -1.3246546983718872),
          ('애견동반', 0.453514039516449, 10, -2.430345296859741),
          ('일반칼국수', 0.3144122064113617, 12, -1.9776806831359863),
          ('스타일리스트', 0.30995041131973267, 5, -0.03425504267215729),
          ('가스라이팅', 0.3089645802974701, 8, -1.7649964094161987),
          ('메이크업샵', 0.23990167677402496, 4, -1.8633829355239868),
          ('니다남편', 0.19404704868793488, 17, -2.3800368309020996),
          ('응원법', 0.19351451098918915, 8, -1.5448956489562988),
          ('하하하하하', 0.19188380241394043, 5, -1.035581111907959),
          ('요남편', 0.18237920105457306, 37, -1.3809949159622192),
          ('요남자친구', 0.177988201379776, 7, -0.9567921757698059),
          ('코시국', 0.17575977742671967, 8, -0.20036575198173523),
          ('첫번째', 0.1731429398059845, 9, -2.5559778213500977),
          ('습니다남자친구', 0.1718582808971405, 5, -0.21648401021957397),
          ('니다저', 0.16151316463947296, 63, -1.3341434001922607),
          ('시외숙모', 0.16132156550884247, 6, -1.0740036964416504),
          ('초등학생때', 0.1582733392715454, 4, -2.9970004558563232),
          ('멤버들', 0.15588442981243134, 18, -1.9795150756835938),
          ('앰플쓰고', 0.15249690413475037, 3, 0.096438214182853 7)]
```

```
kiwi.load_user_dictionary('new.txt')
```

```
4
```

```
inputs = list(open('natepann.txt', encoding='utf-8'))
kiwi.extract_words(inputs, min_cnt=1, max_word_len=10, min_score=0.25, pos_score=-3.0,
```

```
[('서이것저것', 0.4674535095691681, 6, -1.3246546983718872),
 ('일반칼국수', 0.3144122064113617, 12, -1.9776806831359863)]
```

구현 진행 사항

- “단순 비교” 후 “Kiwipiepy” 라이브러리 사용하여 새로운 용어 추출

  → 크롤링하여 형태소 단위로 추출한 텍스트와 사전을 비교하여 사전에 없는 미등록 용어 모두 추출

  → 추출한 미등록 용어에 “Kiwipiepy” 라이브러리를 사용하여 명사 형태의 “새로운 용어“ 추출

  ⇒ 스우파, 스트릿우먼파이터, 재난지원금 등 현재 트렌드를 반영하는 새로운 용어 추출

  ⇒ “Kiwipiepy”만 사용하는 것보다 추출 결과가 유의미함

```
In [5]: kiwi.extract_words(inputs, min_cnt=1, max_word_len=10, min_score=0.25, pos_score=-3.0, lm_filter=True)

Out[5]: [('재난지원금', 0.5066719651222229, 6, -2.988631248474121),
         ('엔시티', 0.4403059780597687, 9, -2.8248293399810 79),
         ('투바투', 0.3939288854598999, 7, -2.16601443290 71045),
         ('스트릿우먼파이터', 0.3177706599235 5347, 3, -2.677033662796 0205),
         ('스우파', 0.2559112012386322, 10, -2.060888528823 8525)]
```

특허 아이디어 구체화

- 추출한 새로운 용어를 5가지로 분류

1. 초성어

   Ex) ㅇㅈ(인정), ㄱㅇㄷ(개이득), ㄹㄷ(레디), ㅇㄱㄹ(어그로) 등

2. 합성어

   Ex) 빵셔틀, 존예보스, 벼락거지, 멍청비용 등

3. 파생어 (접두어)

   Ex) 개-, 댕-, 핵-, 꿀-, 갓-, 네다-

4. 파생어 (접미어)

   Ex) -충, -혐, -세권, -리미엄, -각, -니즘, -피셜, -미새, -등이, -비용, - 탱

5. 야민정음

   Ex) 커엽다(귀엽다), 머통령(대통령), 머박(대박), 띵작(명작) 등

특허 아이디어 구체화

- 초성어

- consonants = ['ㄱ', 'ㄴ', 'ㄷ', 'ㄹ', 'ㅁ', 'ㅂ', 'ㅅ', 'ㅇ', 'ㅈ', 'ㅊ', 'ㅋ', 'ㅌ', 'ㅍ', 'ㅎ', 'ㄲ', 'ㄸ', 'ㅃ', 'ㅆ', 'ㅉ']
- vowels = ['ㅏ', 'ㅑ', 'ㅓ', 'ㅕ', 'ㅗ', 'ㅛ', 'ㅜ', 'ㅠ', 'ㅡ', 'ㅣ', 'ㅐ', 'ㅒ', 'ㅔ', 'ㅖ', 'ㅘ', 'ㅙ', 'ㅚ', 'ㅝ', 'ㅞ', 'ㅟ', 'ㅢ']

- 자음과 모음을 리스트에 입력해 두고 추출한 새로운 용어에 자음, 모음이 포함되면 초성어로 분류

- 극성값 부여 방식: 초성어는 문맥을 알아야 이해에 도움이 되므로, FastText를 적용하여 연관 단어들의 극성값 통계와, 해당 초성어가 포함된 예시 문장 제공

특허 아이디어 구체화

- 합성어

- 새로운 용어에서 가능한 조합 토큰화하여 모두 추출

  ex) 빵셔틀: 빵+셔틀, 빵셔+틀

- 사전에 등록되지 않은 토큰 제외

  ex) 빵셔 + 틀의 경우 빵셔가 미등록어로 제외됨

- 쪼개진 토큰 둘 다 사전에 포함되면 합성어로 분류

- 극성값 부여 방식: 두 개 토큰의 극성값을 KNU 사전에서 탐색 후, 그 값들을 평균 낸 극성값과 해당 합성어가 사용된 예시 문장 제공

특허 아이디어 구체화

- 파생어

- 미리 자주 쓰이는 접사를 리스트에 입력해 두고 해당 접사가 포함되면 접두어 혹은 접미어로 분류

- 파생어(접두어): 개-, 댕-, 핵-, 꿀-, 갓-, 네다-
- 파생어(접미어): -충, -혐, -세권, -리미엄, -각, -니즘, -미새, -등이, -비용, - 탱

- 극성값 부여 방식: FastText를 적용하여 연관 단어들의 극성값 통계와, 해당 파생어가 포함된 예시 문장 제공

특허 아이디어 구체화

- 야민정음&기타

<야민정음>

- 미리 자주 쓰이는 야민정음을 입력해 두고, 해당 내용이 포함되면 야민정음으로 분류 후 원래의 한글 단어로 변환하여 제시

  ex) 댕댕이 ↔ 멍멍이, 커여워 ↔ 귀여워, 머머리 ↔ 대머리

- 극성값 부여 방식: 변환된 원래 한글 단어의 극성값과 해당 단어가 사용된 예시 문장 제공

<기타>

- 위의 카테고리에 분류되지 않은 기타 용어
- 극성값 부여 방식: FastText를 적용하여 연관 단어들의 극성값 통계와, 해당 용어가 포함된 예시 문장 제공

차주 계획

1. 단순비교&Kiwipiepy를 사용한 새로운 용어 추출 보완
2. FastText로 추출한 새로운 용어의 주변 단어 추출 테스트
3. 논문 작성 - 한국멀티미디어학회

감사합니다 :)

빅데이터미네이터

손윤식 교수님 주식회사 닥터송 송환구 멘토님